DOCUMENTATION
INCORPORATED

SPECIALISTS IN BASIC AND APPLIED INFORMATION THEORY

2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

APPLICATION OF THE UNITERM SYSTEM AND THE ASSOCIATION OF IDEAS TO A SPECIAL LIBRARY FILE

TECHNICAL REPORT NO. 10

Prepared under
Contract Nonr-1305(00)

FOR THE OFFICE OF NAVAL RESEARCH

August
1955

APPLICATION OF THE UNITERM SYSTEM AND THE ASSOCIATION OF IDEAS TO A SPECIAL LIBRARY FILE[1]

TECHNICAL REPORT NO. 10

Perhaps the best way to indicate the depth of a detailed analysis possible with a Uniterm System as contrasted with systems that make a card or cards for each item is to compare the 1148 punching positions in the total system presently employed by the Prevention of Deterioration Center with the fact that for the first five hundred reports we indexed, out of the same group of reports indexed by PDC, we generated a Uniterm vocabulary of approximately 3000 Uniterms, including names of specific materials, organisms, chemicals, etc. Our average assignment of Uniterms per item was 25. We were also able to index in a few months more than the average annual load of the Prevention of Deterioriation Center. On the matter of the adequacy of our methods for meeting the reference needs of the Center, the Office of Naval Research recently prepared a series of experimental questions, and by using this method, was able to get satisfactory answers. But it is possible to indicate certain theoretical problems that must attend the use of the Uniterm system, and which we hope to solve by employing a new technique of retrieval by association.

---

[1]This report was prepared as part of a joint paper with Dr. Curtis L. Brown, Prevention of Deterioriation Center, National Research Council, entitled, "Two Approaches to the Retrieval of Information from a Special Library." Dr. Brown's paper, "Application of a Simple Punched Card System to a Special Information Service," and this paper, are to be read before the Division of Chemical Literature, American Chemical Society Meeting, Minneapolis, Minn., September, 1955.

The Uniterm system, like all other systems of retrieval, operates with words. And since this operation is frank and unashamed, the Uniterm system is often compared unfavorably with other systems which presumably use "ideas" or "subjects" or "semantic content" rather than words as instruments of storing and retrieving information. I will grant that in love and empathy there is communication, and information may be sent in a wink and received in the viscera without the mediation of language, but I think that you must likewise grant that all formal communication, all formal storage of information, must communicate and store symbols and not meanings, even though our purpose is always to communicate meanings and semantic content.

It is certainly the beginning of wisdom to recognize that words are slippery things and that we cannot make an easy one to one identification of symbol and meaning. The same symbol can mean different things as with homonyms. But surely this elementary insight is not all that is meant by those who contrast words and meanings in indexing systems. To use "wrench" rather than "spanner" to stand for a certain type of physical object or to use "tube" rather than "valve" for another is certainly to choose a word but it is not to dispense with words and confront the searcher with meanings unmediated by symbols. For let us be quite clear about this -- and I wish I could say it as wittily as Fairthorne did in a recent paper before

the joint AGARD-SLA meeting on Documentation -- communication systems and storage systems deal with marks, noises, shapes, sounds, patterns, etc. and not with the semantic content of these things. Undoubtedly, the miracle of communication is that we can get semantic content from one mind to another by manipulating shapes and noises. Let us not denigrate this miracle by asking chemists and librarians to distrust it and to talk to one another only in the unmediated meanings of love and intuition.

One of the ways to make language more precise and a better conveyor of meanings is to use names rather than descriptions. The distinction between names and descriptions is an important and abstruse topic of modern logic and we need to do little more here than indicate the difference. When I say "x is carnivorous", I am describing "x". When I say "x is a carnivora", I am naming "x". Now a description may be false while a name is never false. If I say "the man named Mortimer Taube is reading this paper", this is an attribution of a symbol which is not significantly true or false. But if I say the man reading this paper is wise or a fool, I hope I can get a significant argument on the truth or falsity of one statement or the other.

Now there are certain sciences in which the activity of assigning names plays a very important role. In these sciences if everything has only one name and these names are used and known universally than we can build orderly communication and storage systems using as symbols only the agreed-upon names. Especially is this true if the names are used in a fixed pattern, e.g., the Cabots being divided

into the Boston Cabots and the New York Cabots. A description can become a name when its attribution to an entity becomes a question of definition and not of fact. Thus, of all systems of communication and storage, taxonomies are the most purely verbal or symbolic and the least concerned with the interplay of symbol and meaning or the semantic content of the symbols.

A language which has been frozen into a taxonomy by the conversion of descriptions into names becomes a mold into which we may force the facts and ideas which are the concern of any science; and it is the usual fate of taxonomies to be discarded en toto when as artificial things they become too far divorced from reality, rather than to grow and change as a living language grows and changes as it attempts to symbolize ever richer conglomerates or systems of facts and experience.

But living languages are slippery things -- they reflect the meanings of days on which they are used and the different people who use them -- how then can we use a living language not only for poetry and song, but for science and the storage of information which is the physical memory of science. Is there a third possibility which does not commit us to the fixed and ultimately futile patterns of a taxonomy or artificial language and yet which lets us make an orderly use of the rich diversity of living language. The answer I think is to be found in what, following Vannevar Bush, we have called retrieval by association rather than by fixed hierarchical patterns. And here once more I must quote another man who has expressed the same thought, independently arrived at, in different words:

"In my opinion, the problem of bringing pertinent vocabularies into coincidence is the crucial one for all documentation systems whether operated by classifications, indexes, or machines. The system must display its vocabulary for selection. For the field of chemistry, this vocabulary is so enormous and is growing so rapidly that it is impossible because of limitations of time and eyesight to read all of the terms for the purpose of making the proper and complete selection. To facilitate selection, indexes give users cross references and display suggestive and new terms in the modifications under headings. The searchers aid in bringing their vocabularies into coincidence by courses of study, textbooks, reference works, reviews, monographs, dictionaries, current literature, indexes, and the like. At present, the process of bringing vocabularies into coincidence for broad, generic searches is probably the slowest of all processes in the use of documentation systems. The important point to note is that all documentation systems, including mechanized ones, are faced with this problem of bringing pertinent vocabularies into coincidence and that if coincidence is not effected then significant information is certain to be missed. Difficulties with broad generic searches, formerly attributed by some of us interested in mechanization solely to indexes, are, in reality, caused by the difficulties of bringing vocabularies into coincidence." [2]

This means that the alternative to a taxonomy for a broad generic search is a mechanized system of bringing vocabularies into coincidence, or as we have put it, displaying the associations in the system.

I have many times told the story of the early days of the Uniterm system when we tried to demonstrate its method by asking visitors to our offices to pick any two cards and find the number common to both.

---

[2] Charles Bernier, "Organizing Abstract Information," unpublished paper read before American Documentation Institute, November 6, 1953.

The cards would be selected and no common numbers would be found. Slightly embarassed, we would suggest another two cards, many times with the same result. To avoid this situation we prepared a "canned sample" of a set of cards having numbers in common, and we gave this set to all casual visitors. What we had discovered, of course, was the lack of suggestiveness or coincidence of vocabulary in a Uniterm system. If you ask a Uniterm system what it has on "A" and "B", it will give you an answer but if you ask the system to suggest terms which define a subject together with "A", the system remains mute and helpless. And this situation of wondering what is the right question to put to the system is more difficult, the richer, the more flexible and more detailed the vocabulary of the system. Think of the situation as it might face the user of a Uniterm system in the PDC -- with its concern for "thousands of distinctly different materials such as alloys, elastomers, and fibrous products, each available in a multitude of forms and formulations. . . [and] an equal if not greater variety of biological organisms and chemical agents which have the unpleasant habit of attacking materials and structures. . . [and] a legion of accepted and experimental pesticides, oxidation and corrosion inhibitors, fire and waterproofing treatments, and other special-purpose preservatives, for which the systematic nomenclature alone is enough to drive an indexer or coder crazy." I do not approach this system as I might if making a patent search, to inquire whether anyone may utilize a certain composition for a certain purpose. Rather, I may want to know all the compositions that have been used for a certain purpose and I'm unable

to think of their names. Or I want all the uses to which a certain composition has been put and I'm unable to think of all the terms which should be searched. In ordinary affairs our memories help us to recall the symbols of meanings we wish to communicate, but when I interrogate a system which is not a creation of my own experience, my memory cannot help me to know the limitations or the richness of the symbols under which information has been stored. The system must have its own memory which it can display to the searcher on demand. Thus, if we are concerned as in the example given with various types of paints that have been tested under certain physical conditions, a physical memory system must associate all types of paints or coatings with a specified physical condition under which they have been tested. Such a memory device is entirely feasible and we have built a working model in which we have entered certain of the terms of our vocabulary.

One thing we have learned during the past year is that it is virtually impossible to make a manual dictionary of associations. The complexity of memory and associative trails grows so rapidly that only a system of mechanical, optical or electronic channels or relays can perform the proper manipulations and present the proper disclosures. We have employed a combination of punched holes and electrical circuits, in what is basically a very simple system. On a card representing any term in the system we will punch a hole for each term associated with the term heading the card. By making electrical contact through such holes we can light up on a screen

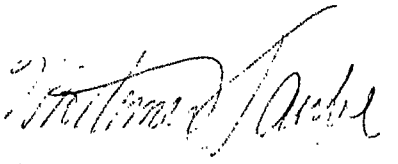
all the words associated with any given word. If from the words displayed we select a term of interest to us we can superimpose the card for this term on the card for the original term. This will black out a certain number of terms on the display which are not associated with both terms of the question, and so on through associations of three, four, and more terms. The more term cards we superimpose on one another, the tighter and more condensed will be our network of associations.

The importance of suggestiveness or a mechanism of associations in an information system varies from field to field. In a theoretical science where connections are theoretically predictable, we do not need such a mechanism. We do not have to be reminded of the "association" of mass and energy. But a memory that tell us that certain paints resist certain organisms and that the system contains positive information on this fact must enhance the retrieval potentialities of the system.

Since this is a symposium on punch cards, one final word is in order on the possible mechanisms which can be used for systems of association. In a previous paper we expressed the conclusion that since an association system was concerned with terms and ideas rather than with items, the Uniterm or Batten system of a card per term or idea was a much more suitable base than a punched card machine employing a card or cards per item. We have since been informed that the more elaborate IBM machines like the 701 can manipulate associations and can in effect superimpose onepattern of associations on another

and point out the common associations. I don't know exactly how this is done, but I would encourage anyone with a 701 machine available to him to give it a try.

Respectfully submitted,



Mortimer Taube
President

August, 1955